DATA PAPER

An Evaluation of Chinese Human-Computer Dialogue Technology

Zixian Feng[1], Caihai Zhu[1], Weinan Zhang[1†], Zhigang Chen[2], Wanxiang Che[1], Minlie Huang[3] & Linlin Li[4]

[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China
[2]AI Research Institute, IFLYTEK CO., LTD., Hefei 230088, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[4]Consumer BG, HUAWEI Technologies CO., LTD., Nanjing 210012, China

Keywords: Chinese human-computer dialogue evaluation; Evaluation data; Few-shot learning; Knowledge-driven multi-turn dialogue

Citation: Feng, Z.X., et al.: An evaluation of Chinese human-computer dialogue technology. Data Intelligence 3(2), 274-286 (2021). doi: 10.1162/dint_a_00090
Received: November 18, 2020; Revised: January 27, 2021; Accepted: February 3, 2021

ABSTRACT

There is a growing interest in developing human-computer dialogue systems which is an important branch in the field of artificial intelligence (AI). However, the evaluation of large-scale Chinese human-computer dialogues is still a challenging task. To attract more attention to dialogue evaluation work, we held the fourth Evaluation of Chinese Human-Computer Dialogue Technology (ECDT). It consists of few-shot learning in spoken language understanding (SLU) (Task 1) and knowledge-driven multi-turn dialogue competition (Task 2), the data sets of which are provided by Harbin Institute of Technology and Tsinghua University. In this paper, we will introduce the evaluation tasks and data sets in detail. Meanwhile, we will also analyze the evaluation results and the existing problems in the evaluation.

1. INTRODUCTION

At the end of the 20th century, with the rapid development of computer technologies, human-computer interaction research came into being [1]. In the 21st century, human-computer interaction research has attracted more and more attention [2,3]. Starting from the Turing test [4], the human-computer dialogue

† Corresponding author: Weinan Zhang (Email: wnzhang@ir.hit.edu.cn; ORCID: 0000-0001-5981-4752).

system has become the research direction of many scholars. Traditional human-computer dialogue systems can be divided into two classes [5,6]. One is the task-oriented dialogue system [7,8] which serves users in accomplishing complex tasks through multi-turn conversations, and the other is the open-domain dialogue system [9,10] which is born purely for small talks. However, the evaluation of large-scale Chinese human-computer dialogues is still challenging.

There are two important tasks in a dialogue system. One refers to few-shot learning in spoken language understanding (SLU). Its purpose is training a model that borrows the prior experience from the old (source) domains and adapts to the new (target) domains quickly even with very few labeled samples (usually one or two samples per class). In recent years, artificial intelligence (AI) has made remarkable achievements with the help of deep learning methods. However, current deep learning methods require a large amount of labeled training data, and a large amount of manually labeled data is often difficult to obtain [11]. Taking task-oriented dialogues as an example, it is often difficult to obtain real user corpus of functions to be developed during product development. Even with raw corpus, task-oriented dialogue development faces the challenge of the high cost of manual data annotation. At the same time, AI applications such as dialogue systems often face the problem of frequent changes in demand, resulting in heavy data labeling tasks that often need to be repeated. However, human beings only need a few examples when learning a new task. This huge contrast inspires researchers to start exploring AI systems that can, like humans, learn from previous experience and from a small amount of data.

The other is knowledge-driven multi-turn dialogue competition. Its purpose is to generate a dialogue response that conforms to the knowledge graph information and context logic when the context and all the knowledge graph information is known [12].

In short, in order to develop evaluation technologies for human-computer dialogue systems, and to provide a good communication platform for academic researchers and industry practitioners, we held the Evaluation of Chinese Human-Computer Dialogue Technology during the Ninth China National Conference on Social Media Processing[①] (SMP2020-ECDT), which consists of two tasks:

1) **Few-shot Learning in SLU.** This evaluation focuses on few-shot learning where there are only a few labeled examples for each test category. The model is first trained in domains with sufficient data, and then tested in a new domain.
2) **Knowledge-driven multi-turn dialogue competition.** The submitted models need to generate a dialogue response that conforms to the knowledge graph information and context logic when the context and all the knowledge graph information are known.

The knowledge graph is described by a series of triples (such as <head entity, relationship, tail entity>). The generated response needs to be fluent enough, semantically relevant to the dialogue context, and conform to the relevant knowledge graph information.

[①] http://smp2020.cips-smp.org/

Compared with SMP2019[2]-ECDT, this year we provided new data sets [13] for each of the two tasks. We conducted natural language understanding for few-shot in Task 1, and we added knowledge to the dialogue competition.

The rest of the paper is organized as follows. We introduce two tasks in detail in Section 2 and give the data sets of two tasks in Section 3. Part of the evaluation results is given in Section 4 and finally the conclusion is made in Section 5.

2. THE FOURTH EVALUATION OF CHINESE HUMAN-COMPUTER DIALOGUE TECHNOLOGY

In this section, we give a brief introduction to evaluation tasks.

2.1 Task 1: Few-shot Learning in SLU

This evaluation focuses on few-shot learning where only a few labeled examples are available for each test category. The model is first trained in domains with sufficient data, and then tested in a new domain.

We give the model a labeled support set as a reference, and let the model mark any unseen query set with user intentions and slots. Taking the test field in Figure 1 as an example, when given the support set and the query sentence "Play Avatar", the model needs to predict that the intent is "Play movie" and the slot is [movie: Avatar].

Figure 1. An example of intent and slot predicting.

[2] http://smp2019.cips-smp.org/

Many text categorization tasks use F1-score as evaluation metric, such as [14].

For the few-slot filling task, we use F1-score as the evaluation index $F = 2PR/(P + R)$, where the average precision as $P = \frac{1}{N}\sum_{n=1}^{N} P_n$ and the average recall as $R = \frac{1}{N}\sum_{n=1}^{N} R_n$. When a key-value combination of the predicted slot is exactly the same as a key-value combination of the ground truth, it is regarded as a correct prediction.

For the intent recognition task, we use the intent accuracy rate (Intent acc) as evaluation index.

In order to comprehensively consider the capabilities of the model, we finally use the sentence accuracy rate (Sentence acc) to measure the comprehensive ability of intent recognition and semantic slot filling.

We give three separate rankings as a reference, and the final ranking of the competition is subject to Sentence acc.

2.2 Task 2: Knowledge-Driven Multi-Turn Dialogue Competition

Task 2 is described as follows: Knowing the dialogue context and all knowledge graph information, models are required to generate dialogue responses that conform to the knowledge graph information and context logic.

In the preliminary stage, we use automatic metrics to evaluate the submitted systems. We choose the following metrics in Task 2:

BLEU-4 [15]: Evaluate the *n*-gram overlap between the generated response and the ground truth.
Distinct-2 [16]: Assess the diversity of the responses.

We calculate the ranking of each model on the above two indicators separately, and use the average of each indicator's ranking as the basis for the ranking.

In the final stage, the top 10 dialogue systems in the ranking list are selected for manual evaluation. In the manual evaluation process, 100 dialogue samples will be selected from the test sets in the three fields, and the responses generated by each team are evaluated in the following two aspects using crowdsourcing:

Informativeness: The amount of relevant knowledge graph information that generated responses contains 3 integers from 0 to 2.

Appropriateness: Whether the generated responses conform to people's daily communication habits.

The final ranking is based on manual evaluation results, containing 3 integers from 0 to 2.

3. EVALUATION DATA SET

The data set in Task 1 is FewJoint provided by Harbin Institute of Technology. It contains 59 real domains, which is one of the most domain data sets. It can reflect the difficulty of real natural language process (NLP) tasks, breaking the current limitations of few-shot NLP that can only perform simple man-made tasks such as text classification.

The source of user corpus mainly includes two parts:

1) Corpus from real users of the iFLYTEK AIUI[3] platform; and
2) Corpus artificially constructed by domain experts.

The ratio of the two data sources is approximately 3:7.

After labeling each data record with user intent and semantic slot, we divide all 59 domains into 3 parts: 45 training domains, 5 development domains, and 9 test domains. We reconstruct the test and development domain data into a small sample learning form: each domain contains an artificially constructed *K*-shot support set and a query set composed of other remaining data. Table 1 shows the statistics of the data set in Task 1. The data set in Task 1 is available for reference[4].

Table 1. Statistics of the data set in Task 1.

Item	Count
Utterance	6,694
Average utterance length	9.9
Total domain	59
Train domain	45
Dev domain	5
Test domain	9
Intent	143
Intents per domain	2.42
Slot	205
Slots per domain	3.47

The data set for Task 2 is KdConv, a Chinese multi-domain data set towards multi-turn knowledge-driven conversation that is provided by Tsinghua University. KdConv contains 86K utterances and 4.5K dialogues in three domains including film, music and travel. Each utterance is annotated with relevant knowledge facts in the knowledge graph, which can be used as a supervision for knowledge interaction modeling. Table 2 shows the statistics of the data set in Task 2. The data set in Task 2 is available for reference[5].

[3] https://aiui.xfyun.cn/index-aiui
[4] https://atmahou.github.io/attachments/FewJoint.zip
[5] https://github.com/thu-coai/KdConv

Table 2. Statistics of the data set in Task 2.

Item	Count
Utterance	86K
Dialogue	4.5K
Average turn	19
Domain	3

4. EVALUATION RESULTS

This part shows partial evaluation results of Task 1 and Task 2. At the same time, we conduct a qualitative analysis of the results. The complete leaderboards are shown in Appendix A.

4.1 Task1

For Task 1, we have received eight submitted systems in the test data set, and parts of the results are shown in Table 3.

Table 3. Top 5 teams of Task 1.

Ranking	Participant	Intent acc	Slot F1	Sentence acc
1	AILab-CC, China Merchants Bank	0.8398	0.8043	0.7086
2	SpeechLab, Shanghai Jiao Tong University	0.8430	0.8209	0.6814
3	Peking University	0.8689	0.7523	0.6774
4	MOE Key Laboratory of High Confidence Software Technologies, the Chinese University of Hong Kong	0.8608	0.7481	0.6763
5	ICRC, Harbin Institute of Technology (Shenzhen)	0.8135	0.7246	0.5924

We find that all teams performed well in intent recognition. It may be because intent recognition is a simple classification task while the slot filling task is more complicated. Surprisingly, we find that the second team performed better than the first in Intent acc and F1, but the final result is worse. This may indicate that the first model has stronger joint training capabilities.

4.2 Task 2

Five groups submitted their systems. We have listed the Informativeness scores and Appropriateness scores of the three domains, respectively, and the final score represents the final results and parts of the results are shown in Table 4.

Table 4. Top 3 teams of Task 2.

Ranking	Participant	Appropriateness			Informativeness			Final results
		Film	Music	Travel	Film	Music	Travel	
1	Suzhou KidX.AI Education Technology Co., Ltd.	1.77	1.76	1.88	1.48	1.52	1.80	1.7017
2	Fuxi Lab, NetEase	1.76	1.79	1.89	0.82	0.93	1.34	1.4217
3	Soochow University	1.73	1.78	1.87	0.68	0.92	1.44	1.4033

From the results above, we find that all teams performed well in **Appropriateness** score, and it indicates that people have gradually learned how to make machines more like humans. But most models failed to use the knowledge, only Model 1 performed better in **Informativeness** score and the score is above 1 in the three domains. Meanwhile, all the teams obtained higher scores in the travle domain than others.

4.3 Analysis

The human-machine dialogue evaluation has been successfully concluded. All participating teams have objectively evaluated their models on the data set provided by us. The participating teams can optimize their models in a targeted manner based on the evaluation results.

4.3.1 Task 1

In the Task 1, in order to solve the problem of few-shot data scarcity, the participating teams used pre-training models, such as BERT [17] and ERNIE [18]. Since the pre-training models can learn generalized language information in a large amount of unlabeled text, it is often used as a basic encoder to transform natural language sentences into hidden states. Participant teams are focused on how to use the dependencies between labels [11] or the rules to complete the mapping from support set to test set.

In order to explore the methods of the contestants, we introduce the models of the top three teams in detail and compare the differences between them.

China Merchants Bank AILab-CC. One of the ways to solve data scarcity problem in NLP is data augmentation. For data augmentation of slot tagging, sentence generation based methods are explored to create additional labeled samples. First, AILab-CC used the synonym words to expand the data for slot recognition and balanced the data to help the model learn the information of different slots. Second, with the help of Hou's paper [11], they used Roberta-wwm-ext [19] as a benchmark model, and fine-tuned the model in the support set. Finally, in order to complete the intent recognition task, they adopted the intent information into the slot recognition. For example, the intent is cov_length, the slot is srcLengthUni, and the result is srcLengthUni- cov_length. However, in their experiments, the introduction of intent information actually reduced the effectiveness of the model. In order to achieve better competition results, they trained the Bert+BI-LSTM+CRF [17] model to complete the sequence labeling task and trained Joint-Bert [20] to complete the intent recognition task. They chose the voting method to complete the fusion of each model.

First, all the models were merged, and then the models were removed one by one. If taking a model out reduces the results, keep this model.

Shanghai Jiao Tong University-SpeechLab. They also used BERT as the encoder. Their model used ProtoNet [21] on the basis of Hou's paper [11] to complete the mapping of the support set to the test set, and has achieved desirable results. They used BERT to encode the support set into a hidden state, and converted it into a sentence vector by averaging the word vector, and then merged it with the input x in the form of vector dot product. Finally they completed intent recognition and sequence labeling tasks through softmax or CDT-CRF.

Peking University. Their method is relatively simple. They built a few-shot language understanding model through pre-training models and rules. Specifically, they used ERNIE as a pre-trained language model, and fine-tuned on the support set, and finally used rules to correct it. In terms of data processing, they built a slot dictionary to improve the accuracy.

4.3.2 Task 2

In Task 2, there are three challenges.

- How to model knowledge?
- How to incorporate knowledge information into the model?
- How to ensure that the model selects the correct knowledge among the candidate knowledge?

Most of the teams used encoders to encode knowledge, and then input it into the pre-training model to integrate knowledge and context.

Suzhou KidX.AI Education Technology Co., Ltd. They trained a topic extraction model to extract all the topics related to the knowledge in the context, and established a connection with the knowledge. Then they used the inverted index model to index all knowledge entities. In the generation stage, for each topic word that appeared in the context, they added a corresponding knowledge into the input. They tried three methods to integrate knowledge and context together.

NetEase Fuxi Lab. They stored all knowledge in the knowledge base and used heuristic knowledge of rule intervals. The heuristic rules used include:

- Relation screening: According to the statistics of the triples given in the training set, the commonly used relations are calculated;
- Head entity screening: Consider matching the head entity from the test to the entity that is easily confused with the common words (such as "dao", "yes"), based on the training set matching head entity word frequency statistics of the three types of knowledge base appearance frequency (in dialogue units). In addition, for some numbers, year and date (such as "1998") entity information, which is confusing, it is filtered by regular matching.

- Confusing entity screening: Some header entities are explained with brackets in knowledge base, but the content of brackets will not appear in the dialogue. Other entities appear in parentheses without annotations and parentheses with annotations and refer to different knowledge, such as "Recognize it", and "Recognize it (Eason Chan Album)". When processing, first save a de-parenthesis dictionary, and match in the form of no parentheses. If there is no matching results in the knowledge base, look up the entity annotated with parentheses from the dictionary.

They input the knowledge and context into the encoder for encoding, and used different attention in the decoder to make the output to attend to the context and knowledge, respectively, and finally added them.

Soochow University. They used knowledge encoder and context encoder to encode knowledge and context, respectively, and used the KL loss in KG Fusion to help the model learn how to choose the correct knowledge. The knowledge and context selected by KGFUSION were input to the decoder to learn how the sound field contained knowledge information. At the same time, in order to ensure semantic relevance, they also added reconstruction loss.

Through this evaluation, people began to pay more attention to few-shot learning and knowledge-driven technologies. From the perspective of the proportion of participating teams, we found human-computer dialogue evaluation has attracted the extensive attention of academia and industry.

5. CONCLUSION

We successfully held the fourth Evaluation of Chinese Human-Computer Dialogue Technology. In this paper, we introduced the two tasks of this evaluation, respectively, and explained the corresponding evaluation indicators. In addition, we illustrated the two data sets of the two tasks in detail. Finally, we analyzed the evaluation results. We hope our work will provide some inspiration for the future evaluation of human-machine dialogue research.

ACKNOWLEDGEMENTS

We would like to thank Social Media Processing committee of Chinese Information Processing Society of China (CIPS-SMP) for its strong support of this evaluation. Thanks to Huawei Technologies Co., Ltd. for providing financial support for this evaluation. Thanks to iFLYTEK Co., Ltd. for providing data and evaluation support. Thanks to Kaiyan Zhang and Jiale Zhang for their indispensable support during the evaluation. This paper is supported by the National Natural Science Foundation of China (No. 62076081, No. 61772153 and No. 61936010).

AUTHOR CONTRIBUTIONS

This work was a collaboration between all of the authors. C.H. Zhu (chzhu@ir.hit.edu.cn) drew the whole picture of the evaluation. W.N. Zhang (wnzhang@ir.hit.edu.cn) is the leader of 2020-ECDT. W.X. Che (car@ir.hit.edu.cn), Z.G. Chen (zgchen@iflytek.com), M. L. Huang (aihuang@tsinghua.edu.cn), and L.L. Li (lilinlin@huawei.com) guided the evaluation process and summarized the conclusion part of this paper. Z.X. Feng (zxfeng@ir.hit.edu.cn) summarized the data sets and results of SMP2020-ECDT and drafted the paper. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

DATA AVAILABILITY STATEMENT

All the data are available in the Science Data Bank repository, https://doi.org/10.11922/sciencedb.j00104.00091, under an Attribution 4.0 International (CC BY 4.0).

REFERENCES

[1] Zhang, W.N., et al.: A Chinese intelligent conversational robot. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pp. 13–18 (2017)
[2] Serban, I.V., et al.: A deep reinforcement learning chatbot. arXiv preprint arXiv: 1709.02349v2 (2017)
[3] Zhang, W.N., et al.: The first evaluation of Chinese human-computer dialogue technology. arXiv preprint arXiv:1709.10217v2 (2017)
[4] Turing, A.M. Computing machinery and intelligence. Mind 59(236), 433–460 (1950)
[5] Wang, X.J., Yuan, C.X.: Recent advances on human-computer dialogue. CAAI Transactions on Intelligence Technology 1(4), 303–312 (2016)
[6] Chen, H.S., et al.: A survey on dialogue systems: Recent advances and new frontiers. arXiv preprint arXiv: 1711.01731 (2017)
[7] Zhang, Y.Z., Zhang, W.N., Liu, T.: Survey of evaluation methods for dialogue systems (in Chinese). SCIENTIA SINICA Informationis 47(8), 953–966 (2017)
[8] Mesnil, G., et al.: Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio Speech Language Processing 23(3), 530–539 (2015)
[9] Yan, R., Zhao, D.Y.: Coupled context modeling for deep chit-chat: Towards conversations between human and computer. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '18), pp. 2574–2583 (2018)
[10] Zhang, W.N., et al.: Neural personalized response generation as domain adaptation. World Wide Web 22, 1427–1446 (2019)
[11] Hou, Y.: Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. arXiv preprint arXiv:2006.05702 (2020)
[12] Zhou, H., et al.: KdConv: A Chinese multi-domain dialogue data set towards multi-turn knowledge-driven conversation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7098–7108 (2020)
[13] Feng, Z.X., et al.: Chinese human-computer dialogue technology dataset. Available at: https://doi.org/10.11922/sciencedb.j00104.00091. Accessed 5 February 2021

[14] Tang, B., Kay, S., He, H.B.: Toward optimal feature selection in NaiveBayes for text categorization. arXiv preprint arXiv: 1602.02850 (2016)
[15] Li, J., et al.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
[16] Papineni, K., et al. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
[17] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805v2 (2018)
[18] Sun, Y., et al.: ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223 (2019)
[19] Cui, Y., et al.: Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv:1906.08101 (2019)
[20] Chen, Q., Zhuo, Z., Wang, W.: BERT for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019)
[21] Zhu, S., et al.: Vector projection network for few-shot slot tagging in natural language understanding. arXiv preprint arXiv:2009.09568 (2020)

APPENDIX A: COMPETE LEADERBOARD

Table A1. The complete leaderboard of Task 1.

Ranking	Participant	Intent acc	Slot F1	Sentence acc
1	AILab-CC, China Merchants Bank	0.8398	0.8043	0.7086
2	SpeechLab, Shanghai Jiao Tong University	0.8430	0.8209	0.6814
3	Peking University	0.8689	0.7523	0.6774
4	MOE Key Laboratory of High Confidence Software Technologies, The Chinese University of Hong Kong	0.8608	0.7481	0.6763
5	ICRC, Harbin Institute of Technology (Shenzhen)	0.8135	0.7246	0.5924
6	Laiye Networktechnology Co., Ltd.	0.7930	0.6944	0.5038
7	1STEP.AI	0.8341	0.5968	0.4752
8	Spoken Dialogue System Lab, South China Agricultural University	0.7784	0.5418	0.4583

Table A2. The complete leaderboard of Task 2.

Ranking	Participant	Appropriateness			Informativeness			Final results
		Film	Music	Travel	Film	Music	Travel	
1	Suzhou KidX.AI Education Technology Co., Ltd.	1.77	1.76	1.88	1.48	1.52	1.80	1.7017
2	NetEase Fuxi Lab	1.76	1.79	1.89	0.82	0.93	1.34	1.4217
3	Soochow University	1.73	1.78	1.87	0.68	0.92	1.44	1.4033
4	Laiye Networktechnology Co., Ltd.	1.62	1.88	1.7	0.47	0.63	0.38	1.1133
5	Ping An Life Insurance Company of China	1.24	1.21	1.09	0.98	0.95	0.9	1.0617
6	TMG, Harbin Institute of Technology (Shenzhen)	1.53	1.53	1.61	0.15	0.25	0.44	0.9183

AUTHOR BIOGRAPHY

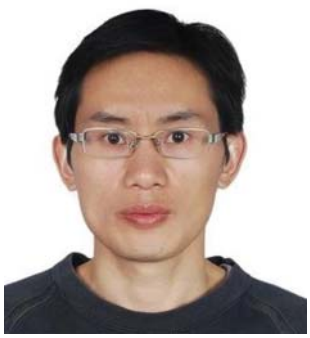
Zixian Feng is a postgraduate in Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. Her current researh interests are mainly in human-computer dialogue system evaluation.
ORCID: 0000-0002-6337-7126

Caihai Zhu is a postgraduate in Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His current researh interests are mainly in conversational recommendation.
ORCID: 0000-0003-3714-1512

Weinan Zhang is an associate professor in Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interest includes human-computer dialogue, natural language processing and information retrieval.
ORCID: 0000-0001-5981-4752

Zhigang Chen joined iFLYTEK Corporation in 2003 and is currently Vice President of the AI Research Institute of iFLYTEK Corporation. He is mainly responsible for cognitive intelligence research and productization.

Wanxiang Che is a professor of School of Computer Science and Technology at Harbin Institute of Technology (HIT). His main research area lies in natural language processing (NLP). He currently leads research projects sponsored by the National Natural Science Foundation of China and the National 973 Project.
ORCID: 0000-0002-3907-0335



Minlie Huang is an associate professor at the Department of Computer Science and Technology, Tsinghua University. His research interests include artificial intelligence, deep learning, reinforcement learning, and natural language processing.
ORCID: 0000-0001-7111-1849

Linlin Li is the leader of Intelligent Voice Assistant, Huawei Consumer BG. She is in charge of Huawei Xiaoyi's voice assistant business. The main work involves the collaboration and mutual assistance of pan-terminal equipment for voice services, multi-language internationalization, multi-modal semantic understanding and end-to-end intelligence.